DOCUMENT RESUME

ED 267 463 CS 505 228

AUTHOR Nadler, Lawrence B.
TITLE Instructor Evaluations in Basic Speech Courses: The Traditional Research Paradigm as a Framework for Analysis.
PUB DATE [83]
NOTE 29p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Course Evaluation; Evaluation Criteria; *Evaluation Methods; Evaluation Needs; Higher Education; Performance Factors; *Speech Communication; *Teacher Effectiveness; *Teacher Evaluation; Test Reliability; Test Validity

ABSTRACT

Noting that instructor and course evaluations represent important, ongoing concerns to students, faculty, and administrators in speech communication, this paper uses the experimental approach to consider the purposes of evaluations, explores major evaluation issues, reviews the literature concerning specific variables possibly related to these ratings, identifies some problem areas, and suggests solutions for the problems. Following an introduction, the paper explores the purposes and methods of evaluation, methodological issues concerning teacher effectiveness assessment, and artifacts in the teaching effectiveness assessment process--a research review, student attributes, situational factors, course characteristics, instructor attributes, and suggestions. The paper concludes with a five-page bibliography. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED267463

INSTRUCTOR EVALUATIONS IN BASIC SPEECH COURSES: THE TRADITIONAL RESEARCH PARADIGM AS A FRAMEWORK FOR ANALYSIS

by

Lawrence B. Nadler
160F Bachelor Hall
Miami University
Oxford, OH 45056
(513) 529-4722

Submitted for consideration for publication in ERIC.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Lawrence B. Nadler

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS 505 228

ABSTRACT

Instructor and course evaluations represent important, ongoing concerns to students, faculty, and administrators. Interestingly, these evaluations are often treated as valid, with little critical examination of the processes of data gathering and treatment. This article urges users of college level evaluations to assess the results of such assessment procedures from the standpoint of the traditional research paradigm-namely, the experimental method by which scientific studies are judged. Within this framework, the author provides an organizing scheme for the many studies which have been performed in this area and draws conclusions concerning the validity of the evaluation process. Additionally, suggestions for improving the evaluation process are offered.

INSTRUCTOR EVALUATIONS IN BASIC SPEECH COURSES:

THE TRADITIONAL RESEARCH PARADIGM AS A FRAMEWORK FOR ANALYSIS

Introduction

Course and instructor evaluations are employed for many purposes, including assessing teaching effectiveness, making personnel-related decisions, and serving as a source of feedback to enhance future teaching. As these functions are extremely important to students, faculty, and administrators, we must ensure that they are valid and reliable indicators of teaching effectiveness in speech communication courses. In multi-section courses (e.g., the introductory speech communication class), this issue is especially significant for two reasons. First, direct comparisons of instructors are possible and may be given greater weight than instances where such comparisons cannot be made. Further, the multi-section course does not entail random assignment of students to specific classes, thereby introducing potentially serious error into the situation.

Using the traditional experimental paradigm can provide a useful, interesting framework for examining course and instructor ratings. Its emphasis on validity and reliability of data is central in examining the worth of instructor evaluations. Thus, this article will utilize the experimental approach to consider the purposes of evaluations, explore major evaluation issues, review the literature concerning specific variables possibly related to these ratings, identify problem areas, and suggest solutions for these problems. While the paper is primarily concerned with the introductory, multi-section course, it will encompass all communication courses.

Purposes and Methods of Evaluation

Before examining specific methods for assessing teaching effectiveness, it is necessary to consider the functions which this process serves. In this regard, Doyle (1983) identifies four such purposes, including the improvement of teaching, acting as an aid in administrative decisions, helping students to choose courses, and providing a criterion for research on teaching itself. Centra (1979) cites similar functions in noting that Deans are utilizing systematic student ratings more than they did in 1973. Given the importance of these purposes, valid and reliable indicators of teaching effectiveness are needed to draw reasonable conclusions. A fruitful starting point entails developing a workable definition of teaching effectiveness. Locksley (1967) offers an interesting perspective by contending that evaluated teaching performance is the sum of teaching ability, environmental influences on teaching, bias toward the instructor, bias toward the field of study, and random errors in teaching performance and observation. From this standpoint, the researcher's task involves developing items which tap into salient teaching dimensions while reflecting a sensitivity to key differences across instructors' styles, courses, and educational settings.

In reviewing 21 studies, Wotruba and Wright (1975) identify the following ten most named traits: communication skills, favorable attitudes toward students, knowledge of subject, good organization of subject matter and course, enthusiasm for subject, fairness in exams, fairness in grading, willingness to experiment, encouragement of students to think for themselves, and interesting lectures. Further, Centra (1973) performed factor analytic studies to delineate distinct, relevant dimensions of teaching effectiveness. He reports that seven factors emerged, encompassing

1. organization/structure/clarity
2. teacher-student interaction/rapport
3. teaching skill/communication/lecturing ability
4. workload/course difficulty
5. grading examinations
6. impact on students
7. global/overall ratings

While factor analytic approaches identify specific dimensions, some educational specialists contend that global ratings are more useful. For example, Centra (1978) maintains that "global ratings of teacher effectiveness or course value (including student estimates of how much they learned) may be more valid estimates of student achievement because they are not tied to a specific instructional style" (p. 40). After determining appropriate dimensions of teaching effectiveness, it is possible to examine various methods for measuring this construct.

The most used and widely accepted method of teaching effectiveness involves student ratings. As Lehman (1974) notes,

> If we subscribe to the definition of a good teacher as one from whom a student learns that which he should learn and define good teaching as the process used to provide maximum learning, the student who sits in a particular teacher's class for the course should be in the best position to provide <u>certain</u> kinds of information" (p. 61).

Centra (1978) supports this justification, observing that "only students, for example, can indicate whether they found the professor to be enthusiastic and stimulating, whether material had been clearly presented, and whether they were challenged by courses in the department" (p. 31). Clearly, student ratings possess face validity and represent meaningful feedback to the instructor. While many criticisms

have been made regarding their validity, Aleamoni (1981) states that the notion that most student rating schemes are simply a popularity contest is false. Further, Costin, Greenough, and Menges (1971) assert that most teachers are receptive to student feedback and are helped by this information. They also note, however, that the belief in low validity of teacher rating scales has produced a decline in their utilization. Other researchers have also identified possible shortcomings of student feedback. For instance, although Centra believes that student ratings are generally valid because they are related usually to student academic achievement (i.e., students learn more from highly rated teachers), he acknowledges that they have certain limitations. Specifically, Centra (1978) posits that "the academic wisdom and experiences of students are limited, and their academic views can be shortsighted" (p. 32). As a result, other approaches to teacher evaluation warrant consideration.

One alternative approach involves using student learning as a criterion for teaching effectiveness. As noted above, this method assumes that more effective teachers have a greater impact on students' learning. In discussing support for this approach, Menges (1977) reports that "a positive rather than a negative relationship is clearly documented by the evidence" (p. 68). Further, some theorists have suggested specific procedures for implementing this rating method and/or using it to support traditional measures of teaching effectiveness. For example, Costin, Greenough, and Menges describe procedures for a multi-section course. They state that "a desirable criterion for validity of a student rating instrument is the rating of a number of instructors whose students are evaluated with the same objective examination. This would seem quite reasonable in larger institutions, where several sections of some undergraduate courses are offered concurrently" (1971, p. 529). This approach would be especially

well-suited to the typical multi-section, basic speech communication course.

Some researchers have cautioned about utilizing student learning as the criterion of teaching effectiveness. Doyle (1983) maintains that it is unfair to evaluate a person in terms of what other people do or fail to do, at least when those people are at best only partially under control of the person being evaluated. Further, he contends that students must be rendered equivalent via statistical correction for student performance to serve as a meaningful criterion of teaching effectiveness. Doyle notes that this residualization method, without random assignment of students to sections, possesses limited utility. Typically, the introductory level speech course is characterized by the absence of random assignment. This criticism, which severely limits the utility of student learning as a criterion of teaching effectiveness, will be examined further in later sections of this article.

Peer and self-evaluations represent two alternative approaches to assessing teaching effectiveness. Centra (1979) asserts that self-evaluations are not especially valid, although they can assist with instructional improvement. He also reports that self-ratings are often more lenient than student ratings (1974). Similarly, Steele, House, and Kerins (1971) observe that teachers are less accurate reporters of effectiveness than their students. Research findings concerning peer evaluations offer little support for the value of this assessment method. For instance, Centra (1975) found that for first year instructors, colleague ratings were substantially more favorable than student ratings. Also, Doyle (1983) summarized this issue by maintaining that "the degree of leniency in student evaluations does not seem to be especially great, but the degree of leniency in colleague evaluations...may be such that the usefulness of colleague evaluations

is severely limited, not only because the ratings are inflated, but also because differentiation among instructors is reduced" (p. 75).

Another approach to assessing teaching effectiveness entails measuring a faculty member's research production. The assumption that being productive in performing research is related to teaching ability underlies this method. According to Costin, Greenough, and Menges (1971), mixed results exist concerning this relationship. Another data source involves alumni ratings. This method presupposes that after students have had an opportunity to observe and experience professional applications of their college classwork, they are in a better position to assess their courses and instructors. Centra (1979) indicates that little difference exists between student and alumni ratings. Finally, some educational specialists have argued that combining several methods might yield the most valid and reliable indication of teaching effectiveness. Centra (1979) posits that "fair personnel decisions can best be made by combining information from several sources so that the shortcomings of one approach can be balanced by the strengths of another" (p. 3). While a multi-method approach is advocated for many purposes, the fundamental issue of whether each teaching effectiveness assessment approach is valid and reliable must be considered. Unless we can demonstrate that any or all of these methods are reasonable indicators of teaching effectiveness, using them alone or in combination with other techniques will make little difference. Before considering the validity of any of these methods (particularly student ratings), it is necessary to explore these important methodological issues in greater detail.

Methodological Issues Concerning Teaching Effectiveness Assessment

In evaluating the results of any data gathering activity, the questions of validity and reliability must be addressed. As noted

earlier, when casting teaching effectiveness ratings in a methodological framework, we must determine whether these ratings can and should be employed for important purposes such as the improvement of teaching and administrative decisions. In most experimental endeavors, random assignment of subjects to experimental conditions is the primary method for promoting internal validity. Without random assignment, alternative explanations for obtained effects (e.g., prior differences between subjects in various experimental conditions) can be readily advanced. As students typically are not randomly assigned to courses or to sections within courses, a methodological flaw in the teaching effectiveness assessment process is evident. In considering the results of different courses, the absence of random selection and assignment may mediate against the validity of comparing teacher ratings from these courses. In a multi-section course, such as the basic speech communication course, the absence of randomization introduces unknown error variance into the teaching effectiveness equation. The importance of random assignment transcends student ratings and is clearly observed in criterion-related studies of teaching effectiveness assessment. For example, Centra (1979) notes that "in short, randomization of students is one of the steps needed to draw a cause and effect relationship between rated teacher effectiveness and student learning" (p. 3). Many variables which may confound student ratings will be explored in the next section.

Random selection of students in college courses also influences the validity of obtained data concerning teaching effectiveness. If students choose some courses based on favorable predispositions toward the topic, instructor, or meeting time, yet select other courses because they are required, then the process of comparing teacher ratings between these courses is more tenuous than when data is gathered using random

selection procedures. Further, students' prior non-random experiences could shape course expectations, which in turn might influence teaching evaluations. While random assignment is not likely to present a major problem in the multi-section basic speech course (unless the combined absence of random selection and assignment produces a situation in which some sections have more majors or otherwise favorably predisposed students than other sections), the danger of comparing course and instructor evaluations in the basic course with other communication classes is readily apparent. Several factors that relate to random selection procedures which may influence student ratings will also be examined in the next section.

Other methodological issues pertain to the rating scales and questionnaire items used to assess teaching effectiveness. Often, the scales themselves are neither balanced nor bipolar in nature. For example, a scale used at one midwestern university had the following scalar points: Below Average-Average-Better Than Average-Exceptional-The Best I Have Ever Had. The lack of equal appearing intervals (or a procedure specifying scale values) prevents straightforward interpretation of obtained ratings, thereby limiting the value of this information. In considering rating scales and questionnaire items, Aleamoni (1981) asserts that validity is promoted by "carefully constructing the instrument so that it contains items and subscales that will yield measures in the areas that are considered appropriate by an individual or group of experts in the field under consideration" (p. 132). In many instances, course and instructor evaluation items are quite general and their validity often is not demonstrated. This issue is especially important in situations where one or two items (i.e., rating the course and rating the instructor) are examined to the exclusion of all other items. As Menges (1973)

observes, in many instances "rating instruments which exhibit adequate stability, internal consistency, and face validity have been used without additional validity research" (p. 66). Thus, the evaluation questionnaire itself is a potentially limiting factor in terms of the obtained evaluation data.

Attrition rates and experimenter bias are two additional problems in collecting course and instructor ratings. In an experimental situation, attrition is normally dealt with by randomly assigning subjects to conditions. In the basic course, where randomization procedures are perhaps non-existent, the dropout rate of various sections could create unwanted error variance and threaten the validity of obtained ratings. Further, as experimenter bias in a research setting can influence a study's results, a similar phenomenon can occur in the course and instructor evaluation context. The major issue involves the possible absence of standardized procedures for administering the rating forms. The instructor might exhibit certain unique intentional and/or unintentional behaviors which could affect students' course evaluations. Similarly, factors such as the timing of administration, the instruction set, and the class mood could influence the teaching effectiveness assessment process. As a result, it is important to examine relevant research to ascertain further the validity of current course and instructor evaluation procedures.

Artifacts in the Teaching Effectiveness Assessment Process: A Research Review

The concern for validity of teaching effectiveness ratings is manifested by the many studies that have been conducted to determine whether various extraneous factors exert significant influence on these ratings. Centra reports an extremely comprehensive study which utilized a population of over 300,000 students in 16,000 classes at more than 100

colleges. Using the Student Instructional Report, 15,000 students in 9,000 classes were sampled and the correlations between traits and ratings were generally insignificant. Certain factors, such as class size, subject area, and the course in relation to the student's curriculum, were significantly linked to the instructor evaluations and the correlations were sufficiently large for these factors to be considered in interpreting the ratings (Centra, 1979). Thus, a closer examination of these and related variables is warranted.

As many variables have been studied, an organizing framework for exploring the literature is useful. The following four broad categories will constitute this framework: Student Attributes, Situational Factors, Class Characteristics, and Instructor Attributes. Some variables may belong in more than one category and overlap between the categories is likely. The value of this framework, though, is its organization of the massive research base and its ability to assist in determining the validity of student ratings of courses and instructors.

Student Attributes

The most widely studied set of factors in the teaching effectiveness assessment realm involves student attributes. These traits include grade point average and ability level, gender, year in college, curriculum of student, personality factors, expected grade, age, and marital status. Without belaboring the issue, these factors are given the opportunity to play an important role in the evaluation process because students are not randomly assigned to sections of the basic speech course.

Several studies (e.g., Rayder, 1968; Costin, Greenough, and Menges, 1971; Miller, 1972; and Rosenshine, Cohen, and Furst, 1973) have found no significant relationship between grade point average and teaching evaluations. Some researchers (e.g., Remmer, Martin, and Elliott, 1949;

Elliott, 1950), though, have found that if an instructor orients his/her teaching to the better students in the class, then GPA is related to teaching effectiveness ratings. Similarly, Centra and Linn (1976) report that in one of three large classes, GPA differentiated subgroups in regard to course evaluations. Thus, GPA interacts with instructor style to influence the teaching effectiveness assessment process.

A related variable involves the grade which a student expects in the course at the time when instructor evaluations are administered. While a few researchers (e.g., Miller, 1972) report no link between expected grade and teacher ratings, considerable evidence for a positive relationship exists. Rosenshine, Cohen, and Furst (1973), for example, note a weak tendency for students who expect a higher grade to give higher ratings. Further, Stewart and Malpass' (1966) examination of 1,975 students in 54 different courses revealed a similar relationship between expected grades and evaluations. Holmes (1970) performed an interesting study in which fifty percent of students who deserved and expected an A or B received it, while the other fifty percent obtained one grade lower than they anticipated. Students subsequently assessed their instructors and on 11 of 19 items lower evaluations were given when the unexpected lowering of grades occurred. Bausell and Magoon (1972) report that students provide lower ratings if their expected grades are below their cumulative GPA. In summarizing this area, Centra (1979) notes that typical correlations are in the .20 range and that the connection between expected grade and ratings of course value is higher generally than the link between expected grade and teacher rating. Interestingly, he speculates that this correlation may be a partial sign of validity, in that it represents modest evidence that students rate courses higher when they learn more. Of course, this connection hinges on the assumption that the course grade can be equated with the level of

student learning.

Students' gender represents another variable which has been studied by educational specialists. While most studies (e.g., Rosenshine, Cohen, and Furst, 1973; Rayder, 1968; and Aleamoni, 1981) have found no association between gender and course evaluations, some evidence for a positive relationship between these variables exists. For instance, Bendig (1952) notes that females were more critical of male instructors than were male students. Similarly, Walker (1969) observes that female students rated female instructors significantly higher than they rated male instructors, while Aleamoni and Hexner (1980) cite five studies in which female students rate instructors higher on some subscales of instructor evaluation forms than male students.

Two other student-related variables entail year-in-college and where a particular course fits in the student's curriculum. Centra and Linn (1976) report that in one of three classes, college year differentiated subgroups of students. Further, Rosenshine, Cohen, and Furst (1973) found a weak tendency for more experienced students to provide higher ratings. The curriculum variable is quite complex, involving distinctions between major versus non-major and required versus elective courses. Although some research endeavors (e.g., Miller, 1972; Rosenshine, Cohen, and Furst, 1971; and Hildebrand, Wilson, and Dienst, 1971) report no link between required/desired course and instructor ratings, considerable evidence exists for a positive relationship. In Centra's (1979) large study, cited in the beginning of this section, the "course in relation to the student's curriculum" variable was consistently related to teacher evaluations. Further, Centra notes that students give slightly higher ratings to major or elective courses than college requirements. Similarly, Miller (1972) and Granzin and Painter (1973) observe that ratings are higher when

students possess a genuine interest in the subject matter. Several other studies (e.g., Cohen and Humphreys, 1960; Gage, 1961; Gillmore and Brandenburg, 1974; and Pohlman, 1975) support the contention that required courses entail lower instructor evaluations than elective courses. The major versus non-major distinction appears to be less significant as a determinant of course evaluations. Aleamoni (1981) discusses four studies involving no relationship, while Costin, Greenough, and Menges (1971) indicate some evidence exists for a positive relationship between these variables. Overall, then, this dimension is often related to teacher ratings.

Three demographic variables which have been examined include the student's personality, age, and marital status. For the latter two factors, Rosenshine, Cohen, and Furst (1973) report no association with course evaluations. As Costin, Greenough, and Menges (1971) indicate, the student personality variable has produced conflicting evidence. Spaights (1967), examining high versus low student achievement groups, reports that the low achievers are more inclined to view instructors as impersonal, dictatorial, sarcastic, and lacking in enthusiasm. McKeachie (1978) uncovered an interaction between student flexibility and teacher authoritarianism which influenced instructor ratings. Similarly, Domino (1971) found an interction between degree of student conformity and teaching approach.

Overall, student attributes appear to be related substantially to course and instructor evaluations. While the research findings are often contradictory, sufficient evidence for this relationship exists in the literature. In the absence of randomization procedures, these factors militate against the validity of teaching effectiveness ratings. In the multi-section, basic communication course, this situation renders comparison of various instructors' ratings tenuous. Further, other sets

of variables must be considered to obtain a more complete picture of the nature of these evaluations. The next section addresses situational aspects of the educational milieu.

Situational Factors

The role of context must be explored in analyzing any communication event. In the course and instructor evaluation realm, situational factors encompass two broad areas. First, enduring situational features, such as class size and the extent to which the class fits into the student's schedule, can be identified. Also, characteristics of the setting in which the rating forms are administered warrant attention. This subset incorporates the nature of the instruction set, the class mood, the timing of the evaluation process, and whether students possess knowledge concerning the purposes or uses of the data.

The issue of class size has received much empirical attention. In Centra's (1979) large study, class size was one of three factors which exerted significant influence on teacher ratings. He reports that classes with less than fifteen students are most highly rated, followed by classes with 16 to 35 students, with classes of 35 to 100 students receiving the lowest evaluations. Further, Miller (1972) asserts that an inverse relationship exists between class size and instructor evaluations. More recently, Aleamoni (1981) has summarized research in this area by noting that eight studies show an inverse relationship, six studies entail a curvilinear association, and seven studies reveal no connection between these variables. A closely related issue involves the number or percentage of students within a class who complete the evaluation forms. According to Miller (1972), "if the percentage of returns for any one class is below 70, the results may be of questionable validity" (p. 30). Also, some educational specialists advise against using ratings from a single class to judge an

instructor's capabilities. For example, Gilmore, Kane, and Naccarato (1978) argue that if only one or two courses are considered in assessing a teacher, the results should not be utilized as a measure of teaching effectiveness for personnel decisions. They maintain that when five or more courses are examined, a class size of 15 students yields reliable data. If ten or more classes are analyzed, class size is a relatively trivial consideration.

Another enduring situational factor involves the time of day when the course is taught and the manner in which the class fits into the student's schedule. Miller (1972) found no relationship between whether a course fits well into the student's schedule and instructor ratings. Further, Aleamoni (1981) reports that time of day has limited impact and is not a significant determinant of student ratings. Hence, this set of factors is unrelated to the course and instructor evaluation process.

The classroom mood during the administration of the rating forms has been examined in terms of its effect on the evaluation process. Costin (1968) required students to assess their courses twice, with the sessions separated by an interval of five weeks. He obtained a correlation of .70 and concluded that mood exerted minimal impact on the ratings. The timing of the evaluation procedures was studied by Costin, Greenough, and Menges (1971), who report that the ratings are stable enough not to be influenced by the quality of the lecture immediately preceding data collection. The accompanying instruction set is another factor which could mediate the evaluation process. Aleamoni and Hexner (1973) contend that the ratings are susceptible to instructional sets, such that telling students that the results will be used for salary and promotion decisions inflates the ratings. Further, Fentress and Swanson (1973) found that the ratings are affected when oral instructions provide reinforcement and praise to the class. This issue is quite

serious, as a few off-hand remarks at the conclusion of the course can influence evaluations which have been formed over a ten to fifteen week period. Also, a lack of control across sections of the basic communication course might mediate against comparisons among instructors. A possible solution to this methodological problem is to have another faculty member administer the evaluations. As Kirchner (1969) found that the presence of the instructor versus a neutral individual produced higher evaluations, this alternate procedure is especially useful.

It is apparent that some situational factors significantly influence the evaluation process. Although conflicting evidence exists in some cases, many studies indicate that the course and instructor evaluation process is plagued by methodological artifacts. Before drawing conclusions regarding the basic communication course, other factors merit consideration. The next section examines the impact of course characteristics on teaching effectiveness ratings.

Course Characteristics

A distinguishing characteristic of any college class is the nature of the course itself. Specifically, classes vary in terms of subject area, level, and format. This section examines the research literature concerning these variables.

In Centra's large study, subject area was one of three variables which was significantly related to teaching effectiveness ratings. Further, he notes that these ratings were slightly higher in the humanities than in the natural or social sciences. Centra (1979) also observes that only in the humanities did faculty see teaching as more important than research.

According to Gage (1961), teachers of lower level courses generally receive lower ratings. Miller (1972) also reports a positive

relationship between course level and evaluations, while Costin, Greenough, and Menges (1971) indicate that graduate classes typically entail higher ratings than undergraduate courses. In reviewing 26 studies concerning this issue, Aleamoni and Hexner (1980) concluded that 18 studies found a positive link between this variable and teacher ratings. Clearly, the comparability of ratings between lower level and advanced courses is quite limited. Although the comparability issue is not problematic within the basic speech course, comparisons between this course and upper level classes may be inappropriate.

A third course factor involves class format. One major distinction entails lecture versus discussion classes. According to Centra (1979), discussion classes generally yield higher ratings than lecture classes. Another concern is the extent to which an instructor serves an entertainment function. Costin, Greenough, and Menges (1971) report no effect for entertainment on course evaluations, but Naftulin, Ware, and Donnelly (1973) found that an instructor with good delivery and weak content was perceived more favorably than an instructor with a weak delivery, regardless of content quality. Thus, they conclude that students are not good judges of a teacher's subject knowledge.

Generally, course characteristics are moderately related to course and instructor evaluations. Some factors are invariant and beyond the instructor's control, while other aspects can be manipulated by the instructor. Further, these factors militate primarily against comparisons between the basic speech course and other classes, although variations across sections of the basic course might also affect teacher ratings. As these latter features are related to individual teachers' choices, it is useful to explore instructor attributes as a category of relevant variables.

Instructor Attributes

The instructor obviously should have an effect on teaching effectiveness ratings. A distinction must be made, though, between factors relating to an instructor's classroom performance and extraneous instructor attributes. This latter category, which includes teacher characteristics, personality, and reputation, is examined below.

Centra (1979) reports that teacher characteristics, such as academic rank and gender, are unrelated to student ratings of teaching effectiveness. When course load exceeds thirteen hours, instructor evaluations are significantly higher than under conditions of a lighter teaching schedule. Obviously, this teaching load is characteristic of the junior and community colleges, where teaching is the main responsibility and research is comparatively deemphasized. In examining teacher personality, Costin, Greenough, and Menges (1971) report no demonstrated link between ratings and instructor personality, suggesting these findings may be due to methodological problems of these studies. As indicated earlier, teacher authoritarianism and approach interact with student characteristics to influence course and instructor evaluations. Finally, a professor's reputation can affect his/her ratings. According to Miller (1972), when students hear that a professor is good, higher ratings are given to that faculty member. Costin, Greenough, and Menges (1971), though, caution against overstating the role of "student gossip" in the evaluation process. They assert that students are generally unaware of how other students rate their instructors. Overall, extraneous teacher characteristics appear to exert minor influence on course and instructor evaluations.

In summarizing research findings in the four categories above, we can note that the impact of these variables ranges from negligible to potent. A key point, in my opinion, is that the combined effect of

these factors must be considered rather than treating each characteristic in isolation. In this light, considerable doubt is cast upon the validity of teaching effectiveness ratings. If the evaluation process is viewed as an application of the scientific method (with the same rigor we apply to research in our communication journals), these artifacts would be sufficient for most researchers to dismiss the experimental findings. Instead, teacher evaluations are often treated as having far more precision than they probably possess. In fact, Turnbull (1968) identifies this problem as the "micrometer fallacy" and argues that small ratings differences should be treated as trivial. Further, some educational specialists have argued that instructors may intentionally attempt to manipulate the ratings (e.g., by giving higher grades). As a result, Centra (1979) suggests that class GPA and related factors should be partialed out statistically. At any rate, the current course and instructor evaluation system is methodologically flawed and considerable attention should be devoted to enhancing the validity and utility of the important data which this process provides. The following section identifies some suggestions for improving the worth of teaching effectiveness ratings.

Suggestions

The purpose of this paper has not been to dismiss summarily the validity of course and instructor evaluations. Instead, the goal has been to identify serious shortcomings of the current process and to suggest possible remedies for this situation. While random assignment would help control many extraneous variables, this procedure is rarely viable in the multi-section speech communication course. As a result, other suggestions which satisfy the feasibility criterion must be explored.

According to Centra (1979), instructors should be rated across five

or more courses using a fairly short rating form with global items. Further, he contends that teachers should have the option to add items and that standardized administration procedures should be employed. Other educational specialists assert that comparison norms must be established to interpret accurately the results of the evaluation process. For example, Doyle (1983) advocates comparing similar courses and posits that "...the broader the norm group, the less likely it is that the requirements of comparability will have been met" (p. 134). Similarly, Costin, Greenough, and Menges (1971) maintain that "if the ratings are to be used as a part of an evaluation of instructors' ability, then courses should probably be compared with others of the same general classification and level (e.g., advanced, intermediate, introductory, lecture, laboratory, seminar, general requirement, elective, etc.)" (p. 529).

Some researchers have suggested specific procedures for evaluating instructors. Farrar (1969), for instance, believes the critical incident technique, in which incidents would be collected from faculty and students and then students would assess the nature and frequency of these incidents regarding specific instructors, is realistic and valuable. More recently, Purdue University's CAFETERIA system, which entails five required items and additional optional scales selected by the individual instructor, is being viewed as an improvement over previous evaluation methods. According to Doyle (1983), this system is perhaps the most valid questionnaire-based instructional evaluation tool available. It remains to be seen whether this system can overcome the shortcomings of its predecessors. Meanwhile, this paper has attempted to illustrate the great importance of instructor ratings, the potentially significant role which extraneous variables can exert on the measurement of teaching effectiveness, and the genuine need to

demonstrate and/or enhance the validity of current evaluation methods.

REFERENCES

Aleamoni, L. M. Student ratings of instruction. Handbook of Teacher Evaluation, ed. J. Millman. Beverly Hills: Sage Publications, 1981.

Aleamoni, L. M. and Hexner, P. Z. The Effect of Different Sets of Instructions on Student Course and Instructor Evaluation. Urbana: University of Illinois, 1973.

Aleamoni, L. M. and Hexner, P. Z. A review of the research on student evaluation and a report on the effects of different sets of instructions on student course and instructor evaluation. Instructional Science, 1980, 9, 67-84.

Bausell, R. B. and Magoon, J. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement, 1972, 32, 1013-23.

Bendig, W. A. Relation of level of course achievement of students, instructor and course ratings in introductory psychology. Educational and Psychological Measurement, 1952, 13, 437-88.

Centra, J. A. Effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973, 65, 395-401.

Centra, J. A. Two Studies of the Utility of Student Ratings for Improving Teaching. Princeton, New Jersey: Educational Testing Service, 1974.

Centra, J. A. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-37.

Centra, J. A. Using student assessments to improve performance and vitality. Evaluating Faculty Performance and Vitality, ed. W. R. Kirschling. San Francisco: Jossey-Bass, Inc., Publishers, 1978.

Centra, J. A. Determining Faculty Effectiveness. San Francisco: Jossey-Bass Inc., Publishers, 1979.

Centra, J. A. and Linn, R. L. Student points of view in ratings of college instruction. Educational and Psychological Measurement, 1976, 36, 693-703.

Cohen, J. and Humphreys, L. G. Memorandum to Faculty. Unpublished Manuscript. Urbana: University of Illinois, 1960.

Costin, F. A graduate course in the teaching of psychology: Description and evaluation. Journal of Teacher Education, 1968, 19, 425-32.

Costin, F, Greenough, W. T., and Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-35.

Domino, G. Interactive effects of achievement orientation and teaching style on academic achievement. Journal of Educational Psychology, 1971, 62, 427-31.

Doyle, K. O., Jr. Evaluating Teaching. Lexington, Massachusetts: D. C. Heath and Company, 1983.

Elliott, D. H. Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education, 1950, 70, 5-61.

Farrar, W. E. Dimensions of faculty performance as perceived by faculty. Dissertation Abstracts, 1969, 29, 3458.

Fentress, J. H. and Swanson, R. A. The Effect of Instructor Influential Tactics on Their Evaluation by University Students. Unpublished Masters Thesis. Bowling Green State University, 1973.

Gage, N. L. The appraisal of college teaching. Journal of Higher Education, 1961, 32, 17-22.

Gillmore, G. M. and Brandenburg, D. O. Would the Proportion of Students Taking a Class as a Requirement Affect the Student Rating of the Course? Urbana: University of Illinois, 1974.

Gilmore, G.M., Kane, M. T., and Naccarato, R. W. The generalizability of student ratings of instruction: Estimation of teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.

Granzin, K. L. and Painter, J. J. A new explanation for students: Course evaluation tendencies. American Educational Research Journal, 1973, 10, 115-24.

Hillebrand, M, Wilson, R. C., and Dienst, E. R. Evaluating University Teaching. Berkeley: University of California, 1971.

Holmes, D. S. The Effects of Disconfirmed Grade Expectancies on Students' Evaluations of Their Instructor: A Study of Attitude Change. Princeton, New Jersey: Educational Testing Service, 1970.

Kirchner, R. P. A Control Factor in Teacher Evaluation by Students. Lexington, Kentucky: University of Kentucky, 1969.

Lehman, I. Evaluating teaching. The Evaluation of Teaching, ed. W. J. Gephart. Bloomington, Indiana: NSPER, 1974.

Locksley, N. A mathematical look at evaluation of teaching. School Science and Mathematics, 1967, 67, 797-98.

McKeachie, W. J. Teaching Tips. Lexington, Massachusetts: D. C. Heath, 1978.

Menges, R. J. The new reporters: Students rate instruction. Evaluating Learning and Teaching, ed. C. R. Pace. San Francisco: Jossey-Bass, Inc., Publishers, 1973.

Miller, R. I. Evaluating Faculty Performance. San Francisco: Jossey-Bass, Inc., Publishers, 1972.

Naftulin, D. M., Ware, J. E., and Donnelly, F. A. The Doctor Fox lecture: A paradigm of educational seduction. Journal of Medical Education, 1973, 48, 630-35.

Pohlman, J. T. A multivariate analysis of selected class characteristics and student ratings of instruction. Multivariate Behavioral Research, 1975, 10, 81-91.

Rayder, N. F. College student ratings of instructors. The Journal of Experimental Education, 1968, 37, 76-81.

Remmers, H. H., Martin, F. D., and Elliott, D. H. Are students' ratings of instructors related to their grades? Purdue University Studies in Higher Education, 1949, 66, 17-26.

Rosenshine, B., Cohen, A., and Furst, N. Correlates of student performance ratings. Journal of College Student Personnel, 1973, 14, 269-72.

Spaights, E. Students appraise teachers' methods and attitudes. Improving College and University Teaching, 1967, 15, 5-17.

Stewart, C. T. and Malpass, L. F. Estimates of achievement and ratings of instructors. Journal of Educational Research, 1966, 59, 347-50.

Steele, J. M., House, E. R., and Kerins, T. An instrument for assessing instructional climate through low-inference student judgments. American Educational Research Journal, 1971, 8, 447-66.

Turnbull, W. W. Testing scores in perspective. Educational Record, 1968, 59, 291-96.

Walker, B. D. An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers. Dissertation Abstracts, 1969, 29, 3474.

Wotruba, T. R. and Wright, P. L. How to develop a teacher-rating instrument: A research approach. Journal of Higher Education, 1975, 46, 653-63.